Towards a comprehensive corpus of eighteenth-century English print: preparation and preliminary studies

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

In this paper, (1) we describe our ongoing work aimed at improving the Eighteenth-Century Collections Online (ECCO) database's utility as a comprehensive and high-fidelity resource, and (2) we discuss some preliminary experiments based on ECCO.

1 Introduction

It is of interest to have a clean, comprehensive corpus of eighteenth-century literature in order to study linguistic and literary questions (1) on as wide a range of documents as possible - not just on a small collection of canonical or specialized documents - and (2) on as high-fidelity a representation of each document as possible - not just on n-gram counts or other summaries. For example, it is of interest to study literary trends and influences over time, and some such trends are likely (1) to pass through or originate in texts that are no longer considered important, and (2) to be encoded in patterns that may not be captured by generic high-level summaries.

There already exist corpora that partially satisfy this need, but, to our knowledge, all current collections have important shortcomings. (1) Several 18th-century collections exist that target specific subperiods, genres, or topics, e.g., Denison and van Bergen (2007) focus on "English letters on practical subjects, dated 1761–90", Müllenbrock (nd) contains pamphlets from 1710–1713, and Sturder (2003) focuses on newspaper articles. These collections are not comprehensive. (2) ARCHER (Biber et al., 1994a; Biber et al., 1994b; ARCHER Consortium, 2010) aims to be a "representative corpus of historical English registers". It includes texts from 11 genres and 8 periods, including the 18th century, but it is not comprehensive within the 18th century. (3) The Google Books n-gram dataset (Michel et al., 2011; Team, 2010) consists of 1- through 5-gram counts from books scanned and converted to text by Google Books. The counts are collected by year and language. However, Google Books' coverage is not yet comprehensive in the 18th century; further, since only n-grams are available for download, it is not high-fidelity in the sense mentioned above. (4) Finally, the Eighteenth-Century Collections Online (ECCO), Part I (Gale Cenage Learning, 2009) is a comprehensive database of 18th century English literature, and the full text of each work is available. However, as described later in the paper, the collection is not as useful as it could be for research on 18th century English literature and language, due to several irregularities, including imperfect optical character recognition (OCR) and the presence of duplicate and non-English documents in the corpus.

In this workshop paper we describe our ongoing work aimed at improving ECCO's utility as a comprehensive and high-fidelity resource, and we discuss some preliminary studies of 18th century language based on ECCO.

2 Basic information about the collection

Before describing our own work, we give details about the provenance and contents of the ECCO database (Gale Cenage Learning, 2009). The ECCO database, initially released as a searchable online

database in 2003 by the Gale Group, is composed of scanned microfilms made in the 1980s for the purpose of making eighteenth-century books available in libraries worldwide. ("Books" here is used loosely; it includes some pamphlets, periodicals, broadsides and the like; the survival and subsequent cataloging of ephemera is irregular and unpredictable.) The microfilm collection was initially based on the British Library's holding of eighteenth-century printed books, which was cataloged in the English Short Title Catalog beginning in 1977. The catalog has since expanded to include the holdings of 2000 libraries worldwide. It consists mainly of books printed in Great Britain and its colonies between 1701 and 1800. Early in the cataloging process, it was decided that multiple editions of books as well as books initially published before 1701 and republished in the eighteenth century would be included. The ECCO database has continued this practice de facto by including all ESTC microfilmed books; ECCO Part I includes all books printed in Roman fonts that were part of the ESTC listings as of 2002.

Our work is based on an OCR'd version of the corpus made available by the Gale Cenage Group. The corpus consists of 112040 documents spanning 42 gigabytes of plain text, and 149546799 unique terms (after §3's preprocessing), including 94084366 hapax legomena. The term "the" occurs 262543504 times.

We also have access to 250 hand-keyed, diplomatic-quality transcriptions of books published during the 18th century. These texts were created by the Text Coding Partnership (TCP) and the ECCO-TCP partnership (Text Coding Partnership, 2011).

3 Simple preprocessing to correct OCR

In order to correct many of the most common errors introduced due to OCR, we preprocessed the text of each document using the following simple rules:

- Replace " ' d" → "'d", e.g., "reform 'd" → "reform'd".
- Replace "& c" → "&c", e.g., "& c" → "&c".
- Replace "- " → "", e.g., "Spi- rit" → "Spirit".
- Replace "-" → " ", e.g., "He boldly hiccups- but he cannot" → "He boldly hiccups but he cannot".

Its Su- burbs, burbs, . & c. are of ' :vast Extent;':but Cairo irfelf, well examinl'd, as to its just Circum- ference, is not much -bigger thain Paris. It is computed to contain near five millions of ii'habitarits; and in it are reckon'd two thousand Mofquer

its suburbs burbs &c are of vast extentbut cairo irfelf well examinld as to its just circumference is not much bigger thain paris it is computed to contain near five millions of iihabitarits and in it are reckond two thousand mofquer

Figure 1: A typical example of raw text and its preprocessed version.

- Remove all characters other than a-z, A-Z, 0-9, "&", and " ".
- Lowercase everything.

The rules were performed one after another.

A typical example of raw text and its preprocessed version is shown in Figure 1. From this example, one can see that our simple rules improve the text substantially in some respects, but also that a smarter approach to tokenization and other basic cleanup can be more successful. At present, we are working on a model of OCR errors in the ECCO collection, using the hand-keyed ECCO-TCP texts and their matching OCR'd ECCO texts as training data. Ultimately, we would like to create a version of the ECCO collection with OCR errors automatically corrected based on this model.

4 Duplicate detection

The original ECCO corpus has a large number of duplicate documents. Duplicates occur in ECCO primarily for two reasons: (1) a book was published in more than one edition during the eighteenth century; (2) a document was included in ECCO more than once as resources from different libraries were added to the collection. In either case, it is often desirable to exclude duplicates from our study of some phenomenon. For example, if we want to study linguistic trends, then it does not make sense to include documents that were republished 50 years after the time they were originally published.

It is therefore of interest to classify each document in the database as either a duplicate of a document with an earlier publication date, or not. In order to do so, we need to make "duplicate" more precise. One natural approach is to say that two documents are duplicates of each other if they share more

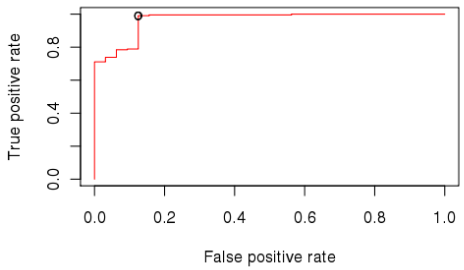

Figure 2: ROC curve for the duplicate detection task. The level corresponding to the threshold $t = 0.35$ is marked with a circle. See text for details.

than a certain percentage of the same words. The following function is ≥ 0.9 if documents d_1 and d_2 share $\geq 90\%$ of their common terms:

$$g(d_1, d_2) = \frac{|\{w \in V : w \in d_1 \text{ and } w \in d_2\}|}{|\{w \in V : w \in d_1 \text{ or } w \in d_2\}|},$$

where $|\cdot|$ denotes cardinality and V is the vocabulary. In other words, g is the Jaccard index applied to the set of terms occurring in each document.

By thresholding the Jaccard index g, we obtain a family of binary classifiers f_t that predict whether a given pair of documents is or is not a pair of duplicates. Specifically, for each possible threshold $t \in [0, 1]$, we define the classifier f_t as $f_t(d_1, d_2) = 1$ if $g(d_1, d_2) > t$ and 0 otherwise.

We evaluate this family of classifiers and choose a suitable threshold t by looking for duplicates of the hand-keyed ECCO-TCP texts in the main ECCO database, since we know that every ECCO-TCP document is a duplicate of at least one ECCO document. For each ECCO-TCP document c, let d_c be the document d in the main ECCO corpus that maximizes the Jaccard index $g(c, d)$. Next, we check by hand whether in fact d_c is a duplicate of c; we denote the ground truth $\ell(c, d_c) = 1$ if so and $\ell(c, d_c) = 0$ otherwise.

Based on these $((c, d_c), \ell(c, d_c))$ pairs, we can now compute the true positive rate and false positive rate of the classifier f_t relative to the ground truth ℓ, for each threshold level t. The ROC curve shown in Figure 2 summarizes this information as t varies.

As the plot shows, our method is effective at identifying most true duplicates without identifying too many false duplicates. Since our goal is to eliminate as many duplicates as possible, without eliminating too many false duplicates, we choose level $t = 0.35$ for future experiments. This level is marked with a circle on the curve. When we apply the duplicate scheme to the whole corpus, using a threshold $t = 0.35$, 33769 documents (30%) are marked as duplicates of earlier documents.

Duplicate detection is a well-studied topic in general. For other approaches, see Rajaraman and Ullman (2010), Chapter 3, and references therein. The approach above works well for us and is already fast enough without using more sophisticated approaches described in the chapter.

5 Language detection

The ECCO collection includes a number of documents that, despite being published in England, are not primarily written in English. The task at a basic level is to classify each document as either English language or not. Many documents in the corpus are partly in English and partly in other languages, to varying degrees: e.g., there are English-language books with non-English quotations, there are foreign-language dictionaries, and there are non-English texts with English facing-page translations.

For some studies, we would like to exclude documents that are not written in English. Thus we want to be able to classify documents as "English" or not. We make the classification task well-posed by saying that a document is "English" if substantially more than the majority of its text - say >75% of the text - is in English.

We solved this classification problem as follows. From each document, we sampled six 150-word contiguous blocks of text, and we sent each block separately to Google Translate's language detector. Google Translate labeled each block as English or some other language, resulting in 6 "votes" per document. We classified each document as English if at least three of its blocks were classified as English. On the entire corpus, 9570 documents out of 112040 total (8.5%) were classified by this procedure as non-English.

In order to access the accuracy of this classification procedure, we did the following. We sampled 250 documents from the ECCO corpus, and we la-

	00s	10s	20s	30s	40s	50s	60s	70s	80s	90s
00s	1.0000	0.9935	0.9917	0.9870	0.9779	0.9763	0.9646	0.9565	0.9484	0.9463
10s		1.0000	0.9923	0.9900	0.9802	0.9759	0.9664	0.9588	0.9528	0.9500
20s			1.0000	0.9958	0.9904	0.9889	0.9802	0.9734	0.9657	0.9628
30s				1.0000	0.9932	0.9905	0.9847	0.9788	0.9718	0.9692
40s					1.0000	0.9962	0.9896	0.9873	0.9814	0.9794
50s						1.0000	0.9935	0.9904	0.9844	0.9834
60s							1.0000	0.9927	0.9914	0.9878
70s								1.0000	0.9942	0.9917
80s									1.0000	0.9958
90s										1.0000

(a)

	00s	10s	20s	30s	40s	50s	60s	70s	80s	90s
00s	1.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
10s		1.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
20s			1.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
30s				1.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
40s					1.0000	0.0015	0.0001	0.0001	0.0001	0.0001
50s						1.0000	0.0001	0.0001	0.0001	0.0001
60s							1.0000	0.0001	0.0001	0.0001
70s								1.0000	0.0002	0.0001
80s									1.0000	0.0001
90s										1.0000

(b)

Table 1: (a) The observed cosines between count vectors, averaged by decade. (b) The levels of the observed cosines against the permuted cosines. Both matrices are symmetric, and only the upper triangles are shown.

votes	#English	#not English	#total
0	0	9	9
1	0	5	5
2	0	3	3
3	2	1	3
4	6	2	8
5	21	1	22
6	199	0	199

Table 2: This table shows the number of documents that received k "English" votes from Google Translate ($0 \leq k \leq 6$), grouped according to whether the documents were labeled by hand as truely "English" or not.

beled each as being substantially more than half English, or not, by hand. We grouped the 250 sampled documents according to how many (0–6) votes were assigned by Google for English.

The result is summarized in Table 2. Every document that received 6 Google English votes had also been labeled by hand as English, while every document that received fewer than 3 English votes from Google had been labeled by hand as non-English.

Numerous other methods exist to classify documents by language; see, for example, Rehurek and Kolkus (2009) and references therein. However, based on the above results, it seems difficult to beat Google's language identifier for our problem.

6 Decade comparison

We have also investigated large-scale vocabulary change in written English across decades in the 18th century. This study was performed as follows.

First, starting from the tokenized text, we excluded documents that were marked as duplicates of previous documents and documents that were marked as not English. Within each document, we discarded all words occurring $<$100 or $>$5000000 times in the corpus. The goal was to exclude, on one hand, OCR and other noise, and on the other hand, stopwords, since both of these could hide true variations, and in a preliminary experiment on a small dataset, we did observe this before we did thresholding. After this thresholding, 662280 words remained in the vocabulary.

We mapped each document to a bag-of-words count vector, we averaged the count vectors within each decade, and we computed the cosine between the averaged vectors for each pair of decades. The result is shown in Table 1(a), where the (i, j) entry is the cosine between decade i's averaged vector and decade j's averaged vector. Based on this table, vocabulary usage does appear to change throughout the century.

In order to assess the significance of the change, we did a permutation test for significance of the change, as follows. For each pair of decades, we (a) collected the documents occurring in one or the other decade, (b) randomly permuted the labels on the documents as being in one or the other decade, (c) averaged the count vectors according to the new decade assignments, and (d) computed the cosine between the averaged vectors. We did this 10,000 times for each pair and counted the number of times the new cosine (under permuted labels) was less than the observed cosine (under the original labels). Call this number r. The fraction $(r + 1)/(10000 + 1)$ gives an estimate of the probability that a cosine under a random permutation of decade labels would be smaller than the observed cosine. We report this fraction, for each pair of decades, in Table 1(b).

This is not an especially surprising result, of course. However, we plan to use this result as a basis for futher study of word usage and spelling changes in 18th-century literary English.

References

ARCHER Consortium. 2010. Archer 3.2. http://www.llc.manchester.ac.uk/research/projects/archer/archer3_2/.

Douglas Biber, Edward Finegan, and Dwight Atkinson. 1994a. Archer and its challenges: Compiling and exploring a representative corpus of historical english registers. In *Udo Fries, Peter Schneider, and Gunnel Tottie (eds.), Creating and using English language corpora. Papers from the 14th International Conference on English Language Research on Computerized Corpora, Zurich 1993, 1-13. Amsterdam: Rodopi*.

Douglas Biber, Edward Finegan, Dwight Atkinson, Ann Beck, Dennis Burges, and Jene Burges. 1994b. The design and analysis of the archer corpus: A progress report [a representative corpus of historical english registers]. In *Merja Kytö, Matti Rissanen, and Susan Wright (eds.), Corpora across the centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, St Catharine's College Cambridge, 25-27 March 1993 (Language and Computers. Studies in Practical Linguistics 11), 3-6. Amsterdam and Atlanta: Rodopi*.

Gale Cenage Learning. 2009. Eighteenth century collections online: Origins and contents. In *Eighteenth Century Collections Online*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Heinz-Joachim Müllenbrock. n.d. The eighteenth century corpus of pamphlets (ECCOP). *Erfurt Electronic Studies in English*. http://webdoc.gwdg.de/edoc/ia/eese/eccop.html.

Anand Rajaraman and Jeffrey Ullman. 2010. *Mining of massive datasets*. http://infolab.stanford.edu/~ullman/mmds.html.

Radim Řehůřek and Milan Kolkus. 2009. Language identification on the web: Extending the dictionary method. 5449:357–368.

Patrick Studer. 2003. Textual structures in eighteenth-century newspapers: A corpus-based study of headlines. *Media and Language Change: Special issue of Journal of Historical Pragmatics*, 4(1).

The Google Books Team. 2010. Google books ngram viewer. http://ngrams.googlelabs.com/datasets.

Text Coding Partnership. 2011. About tcp. http://www.lib.umich.edu/tcp/about/about.html.

Linda van Bergen and David Denison. 2007. A corpus of late eighteenth-century prose. In *Joan C. Beal, Karen P. Corrigan, and Hermann L. Moisl (eds.), Creating and digitizing language corpora, 2 vols, vol. 2, Diachronic databases, 228-46. Basingstoke and New York: Palgrave*. http://www.llc.manchester.ac.uk/subjects/lel/staff/david-denison/corpus-late-18th-century-prose/.